ELSEVIER
Available online at www.sciencedirect.com
SCIENCE DIRECT®
International Journal of Approximate Reasoning
38 (2005) 1–17
INTERNATIONAL JOURNAL OF APPROXIMATE REASONING
www.elsevier.com/locate/ijar

A fuzzy noise-rejection data partitioning algorithm

William W. Melek [a,*], Andrew A. Goldenberg [a], M.R. Emami [b]

[a] *Robotics and Automation Laboratory, Department of Mechanical and Industrial Engineering, University of Toronto, 5 King's College Road, Toronto, Ont., Canada M5S 3G8*

[b] *Institute for Aerospace Studies, University of Toronto, 4925 Dufferin St., Ont., Canada M3H 5T6*

Received 1 March 2003; accepted 1 March 2004
Available online 11 May 2004

Abstract

Fuzzy C-Means (FCM) and hard clustering are the most common tools for data partitioning. However, the presence of noisy observations in the data being partitioned may render these clustering algorithms unreliable. In this paper, we introduce a robust noise-rejection clustering algorithm based on a combination of techniques that treat the FCM pitfalls with an outliers exclusion criterion. Unlike the traditional FCM, the proposed clustering tool provides much efficient data partitioning capabilities in the presence of noise and outliers. At the conclusion of the theoretical development, we validate the effectiveness of the proposed noise-rejection data partitioning tool through various comparison studies with existing noise-rejection clustering approaches in the literature.

1. Introduction

An important step of fuzzy modeling is the fuzzy rule generation. The system data is partitioned into fuzzy clusters. We can define clustering as partitioning of a group of unlabeled data into a number of clusters such that similar data is assigned to one cluster and data that is less similar is assigned to different clusters. Two main approaches to clustering are typically used: (a) Hard clustering [4], and (b) fuzzy

* Corresponding author. Address: Alpha Global IT Inc., AI Division, 1262 Don Mills Road, North York, Ont., Canada M3B 2W7. Tel.: +1-416-449-2166x310.

E-mail addresses: william.melek@utoronto.ca, william@alpha-it.com (W.W. Melek).

0888-613X/$ - see front matter
doi:10.1016/j.ijar.2004.03.002

C-Means clustering (FCM). Hard clustering assigns each data point to one and only one cluster with a membership grade equal to one, assuming fully defined boundaries between clusters. Practically, the boundaries between clusters cannot be clearly defined. As an alternative, the fuzzy C-Means (FCM) clustering algorithms were suggested [5].

In this paper, the robustness of the fuzzy clustering algorithms is examined. A robust fuzzy model should identify and reject the noise from the training set and eliminate its effect during the system identification, parameter adjustment, and tuning. Although hard and fuzzy C-Means clustering algorithms are used in many applications, they are highly sensitive to the presence of noise and outliers. The FCM algorithm uses the sum of squared errors in its objective function. Thus, this clustering method may fail completely in the presence of noise. As an alternative, in recent years, modified versions of the Possibilistic clustering algorithm (PCM) were introduced to handle the noisy data sets [4], i.e., NPCM. The NPCM clustering algorithm is more robust than the original FCM algorithm in the presence of noise because its objective function involves unconstrained weights that decrease with the distance from the cluster centers. This will result in low weights for the noise points and therefore reduces their effect on the data set. However, this algorithm encounters the same problems as the original FCM algorithms in the sense that some parameters must be selected a priori. Our approach to achieve robustness is based on exploiting the enhancements of the NPCM algorithm [4] to overcome the drawbacks of the original FCM algorithm. In Section 2, we review the fuzzy C-Means clustering algorithm. We also outline the original FCM drawbacks. In Section 3, we review the NPCM noise-rejection clustering algorithm [4]. In Section 4, we introduce the proposed robust fuzzy clustering algorithm as a combination of the two above-mentioned clustering approaches. In Section 5, we validate the effectiveness of the proposed data partitioning tool through four different examples. In Section 6, we discuss the conclusions.

2. The fuzzy C-Means clustering algorithm

The FCM is the most commonly used algorithm for data partitioning. Due to the nature of our applications, i.e., multi-input single-output (MISO) systems, a simpler and applicable FCM clustering is adopted [5]. In this approach, the output data is clustered as a single-dimensional output space. The input space fuzzy partitions are then specified by projecting the output clusters on each of the input variables separately.

For a set of unlabeled data $\mathbf{x} = \{x_1, x_2, \ldots, x_N\} \subset R^s$, where N is the number of data points and s is the dimension of each data feature, the fuzzy clustering is the assignment of c number of partition labels to the features in $\boldsymbol{x}$. c-partition of $\boldsymbol{x}$ are sets of $(c \cdot N)$ membership values $\{u_{ik}\}$ that can be arranged as a $(c \times N)$ matrix $U = [u_{ik}]$. The major step in fuzzy clustering is to find the optimum membership matrix U. The FCM uses the weighted within-groups sum of squared errors objective function J_m by defining the following optimization problem [2]:

$$\min_{(U,V)}\left\{J_m(U,V;X)=\sum_{k=1}^{N}\sum_{i=1}^{c}(u_{ik})^m\|x_k-v_i\|_A^2\right\} \tag{1}$$

where $\mathbf{v}=\{v_1,v_2,\dots,v_c\}$ is the vector of the cluster centers. The matrix A specifies the shape of the clusters. For spherical clusters, A is chosen as an identity matrix. The membership matrix is calculated as [3]:

$$u_{ik,t}=\left[\sum_{j=1}^{c}\left(\frac{\|x_k-v_{i,t-1}\|_A}{\|x_k-v_{i,t-1}\|_A}\right)^{\frac{2}{m-1}}\right]^{-1} \tag{2}$$

where t is the iteration number in the iterative optimization, and m is the weighting exponent. The above original FCM clustering algorithm has three major problems:

- In order to get the optimal partition, initial locations of the cluster centers should be assigned. The FCM algorithm always converges to a local extreme of J_m (weighted within-groups sum of squared errors objective function). Different choices of initial cluster centers may lead to a different extrema.
- The scientific basis for the choice of the weight exponent is still not clear.
- The optimum number of clusters in the data is assigned a priori. There should be a criterion to assign the optimal number of clusters.

In Section 4, possible solutions to treat each of the above-mentioned problems are suggested during the introduction of the proposed noise-rejection clustering algorithm.

3. The possibilistic noise-rejection clustering algorithm

In this section, we study the new Possibilistic C-Means algorithm (NPCM) [4]. The NPCM is more robust than the traditional FCM in the presence of noise and outliers. The NPCM robustifies the PCM by forcing finite rejection of noise and outliers in the data set. The Possibilistic C-Means clustering algorithms use the objective function:

$$J(U,V;X)=\sum_{i=1}^{c}\sum_{j=1}^{N}\left(u_{ij}\right)^m d^2(x_j,v_i)+\sum_{i=1}^{c}\upsilon_i\sum_{j=1}^{N}\left(1-u_{ij}\right)^m \tag{3}$$

where $d^2(x_j,v_i)$ is the distance from a feature point x_j to the cluster center v_i, and υ_i is a resolution parameter and it depends on the number of data partitions. The membership matrix $U=\lfloor u_{ij}\rfloor$ is a global minimum for $J(U,V;X)$ when

$$u_{ij}=\frac{1}{1+\left\{\frac{d^2(x_j,v_i)}{\upsilon_i}\right\}^{\frac{1}{m-1}}} \tag{4}$$

The Possibilistic clustering algorithm defines a cutoff distance for finite rejection of noise in the data. For an ideal Gaussian cluster, the cutoff distance is chosen such that 97.5% of the data points are considered as inliers. For an ideal Gaussian cluster with variance υ_i, $\left\{\frac{d^2(x_j,v_i)}{\upsilon_i}\right\}$ has a chi-square distribution χ^2 with degrees of freedom equal to the dimension of each data feature [8]. The resolution parameter is defined as:

$$\upsilon_i = \frac{\underset{x_j \in c_i}{\operatorname{median}}(d^2(x_j, v_i))}{\chi^2_{0.5}} \tag{5}$$

Then, the cutoff distance can be calculated as follows:

$$d^2_{\text{cut}} = \upsilon_i \chi^2_{0.975} \tag{6}$$

Finally, the membership values can be computed using Eq. (4). If $d^2(x_j, v_i)$ is bigger than the cutoff distance, then the point is identified as noise and it takes a membership grade $u_{ij} = 0$. The reason for using χ^2 distribution is based on the assumption that the clusters could follow a Gaussian distribution. For other types of distributions, the corresponding indices should be used.

The Possibilistic clustering algorithms suffer from the same problems of the traditional FCM algorithm discussed in Section 2. Furthermore, by using the NPCM algorithm, the cutoff distance is calculated assuming that 97.5% of the data points are inliers. However, this criterion may not be applicable for most of the real application. In other words, if we assume that the data in each cluster follow a Gaussian distribution, the percentage of inliers may not necessarily be 97.5%. Hence, there is no scientific basis for the choice of the exact percentage of inliers in the data set. Thus, based on the assumption that the data in each cluster follows a Gaussian distribution, we introduce a noise-rejection criterion that merely depends on the data to be partitioned.

4. A robust noise-rejection fuzzy clustering algorithm

In this section, we define a new algorithm that has a noise-rejection capability as well as a defined criterion to assign the cutoff distance from the data. Initially, we choose the optimum number of clusters to make the fuzzy clusters compact and far from each other [6]. As a result, a validity index is introduced for the choice of the optimum number of clusters. In other words, we minimize:

$$S_{cs}(U, V; X) = \sum_{k=1}^{N} \sum_{i=1}^{c} (u_{ik})^m \left(\|x_k - v_i\|^2 - \|v_i - \bar{v}\|^2 \right) \tag{7}$$

where v_i is center of cluster i, and $\bar{v}$ is the fuzzy total mean vector of the data set considering their belonging to each of the clusters. It can be defined as:

$$\bar{v} = \frac{1}{\sum_{i=1}^{c} \sum_{k=1}^{N} (u_{ik})^m} \sum_{i=1}^{c} \sum_{k=1}^{N} (u_{ik})^m x_k \tag{8}$$

For the selection of the weight exponent, it is suggested to choose it far from its both extremes so as to ensure that the cluster validity index shows the optimum number of fuzzy clusters. In [5], a fuzzy total scatter matrix is defined as

$$s_{\mathrm{T}} = \sum_{k=1}^{N} \left(\sum_{i=1}^{c} (u_{ik})^m \right) (x_k - \bar{v})(x_k - \bar{v})^{\mathrm{T}} \tag{9}$$

The trace of the fuzzy total scatter matrix decreases monotonically from a constant value z to zero as m varies from one to infinity. For data partitioning, a suitable value for m is that which gives a value for trace (s_{T}) equal to $z/2$ [7]. The constant value z is defined as:

$$z = \operatorname{trace}\left(\sum_{k=1}^{N} \left[\left(x_k - \frac{1}{N} \sum_{k=1}^{N} x_k \right) \left(x_k - \frac{1}{N} \sum_{k=1}^{N} x_k \right)^{\mathrm{T}} \right] \right) \tag{10}$$

For the choice of the initial cluster centers, an agglomerative hierarchical clustering algorithm (AHC) is suggested as the initial clustering tool [1]. The AHC algorithm puts each of the n data vectors in an individual cluster. Then, by defining a matrix of dissimilarities $D = [d_{ij}]$, the AHC merges two or more of these clusters, getting to a higher level of data partition. The process is repeated to form a sequence of nested clustering in which the number of clusters decreases gradually until the minimum required number of clusters c is reached. In specific terms, we calculate the $(c \times N)$ matrix of dissimilarities $D = \left[d_{ij}\right]$ as the following Euclidean-based distance:

$$d_{ij} = d(X_i, X_j) = \sqrt{\frac{2n_i n_j}{n_i + n_j}} \left\| v_{\mathrm{h}i} - v_{\mathrm{h}j} \right\| \tag{11}$$

where $v_{\mathrm{h}i}$ and $v_{\mathrm{h}j}$ are mean vectors of the hard clusters X_i and X_j, respectively, and n_i (n_j) is the number of data in the hard cluster X_i (X_j). Next, in order to find the data points that are "too far" from all cluster centers, we propose the following index for each data point x_j in the input set considered [9]:

$$W_j = \sum_{i=1}^{c} \left\| x_j - v_{\mathrm{h}i} \right\|_A \tag{12}$$

where $j = 1, 2, \ldots, N$, c is the number of clusters, and N is the number of data. The index W_j is the summation of the distance of the data point x_j to all cluster centers. This gives a measure of how far each data point is from the different cluster centers assigned in the first step of the algorithm. The noise is identified through the data points that have large values of W_j and, therefore a threshold Ω is assigned to trim these outliers from the data set. Such a threshold is assigned from the data and depends on the upper and lower bounds of the input. After choosing the threshold, we compute:

$$z = \frac{\eta_n}{N} \tag{13}$$

where η_n is the number of noise points and N is the total number of data. The percentage of "good" data points, i.e., inliers can then be calculated as:

$$\hat{z} = (1 - z) \tag{14}$$

After identifying the percentage of inliers in the data, we compute the corresponding chi-square data distribution value [8]. Then, we calculate the cutoff distance:

$$u_{\mathrm{FC}}\mathrm{cut}^2 = \upsilon_i \chi^2 \tag{15}$$

where υ_i is a resolution parameter that depends on the number of clusters, and χ^2 is the chi-square value computed in the previous step. By knowing the new cutoff distance, the optimum number of clusters, the degree of fuzziness, and the initial location of the clusters centers, we calculate the membership matrix through Eq. (4) of the NPCM.

5. Case studies

5.1. Case study A

Fig. 1 shows the example of two Gaussian clusters generated with centers at (50,50) and (150,150) respectively, with 80 data points in each cluster [4]. Ninety one uniformly distributed noise points are added to the data set in the range of $20 \leqslant x_1 \leqslant 200$, $20 \leqslant x_2 \leqslant 200$. Fig. 2 shows the clustering results using the fuzzy C-Means clustering algorithm. The suitable weight exponent is selected as $m = 3.5$. Fig. 3 shows the optimum number of clusters obtained using the fuzzy C-Means clustering algorithm. Fig. 4 shows the cutoff distance selection criterion. Some points

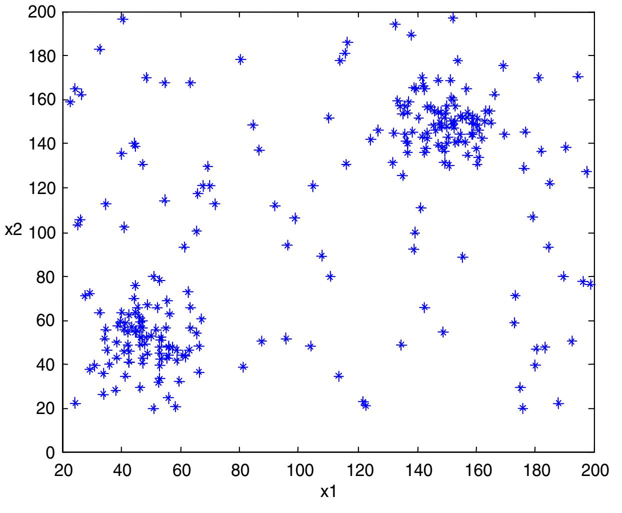

Fig. 1. The Krishnapuram example of two Gaussian clusters.

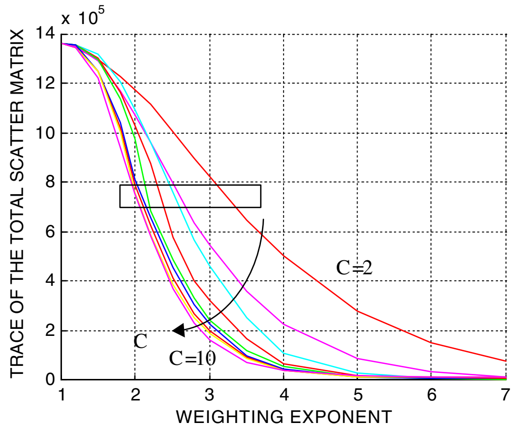

Fig. 2. Selection of level of fuzziness of the fuzzy clustering algorithm for example 5.1.

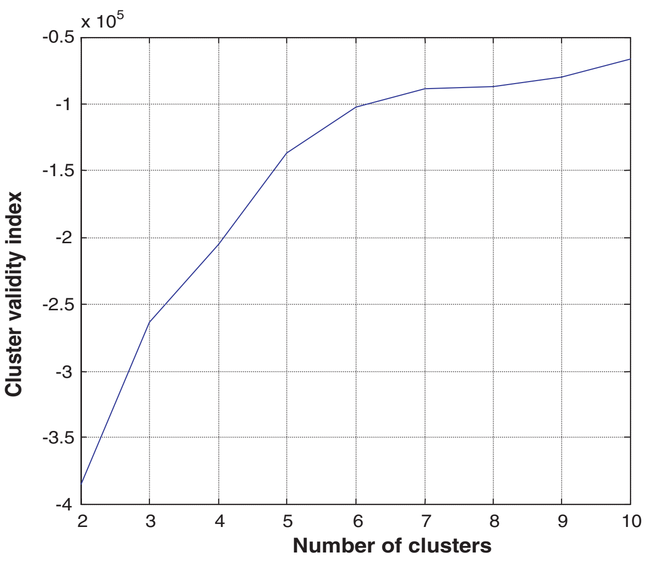

Fig. 3. Identification of he optimum number of clusters for the data in example 5.1.

have large values for the index W. These points basically represent the noise in the data as they are far from all cluster centers, and they are not enough to form a separate cluster. If we choose the cutoff distance at $\Omega = 200$, 22 points on Fig. 4 are rejected. From (13) and (14) we get $z = 0.07$ and $\hat{z} = 0.93$. The cutoff distance is calculated assuming that 93% of the data are inliers. Fig. 5 shows a comparison between the partitions obtained using the NPCM and the new partitions using the improved clustering algorithm. For this example, the noise points are known before performing the comparison between both clustering algorithms. Based on this fact,

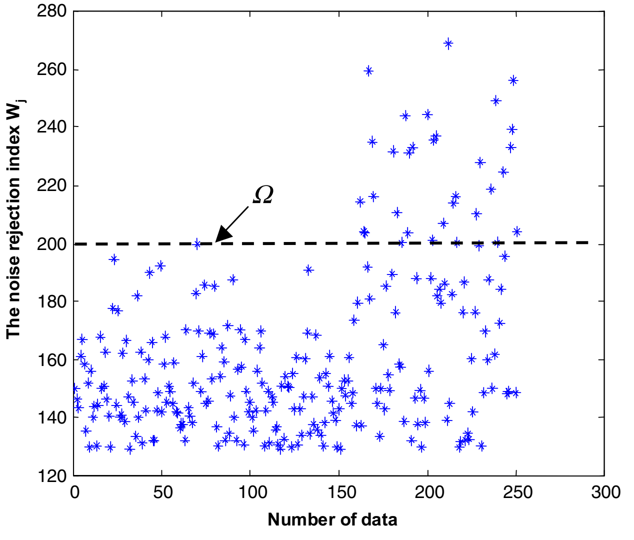

Fig. 4. The application of the noise-rejection criterion for example 5.1.

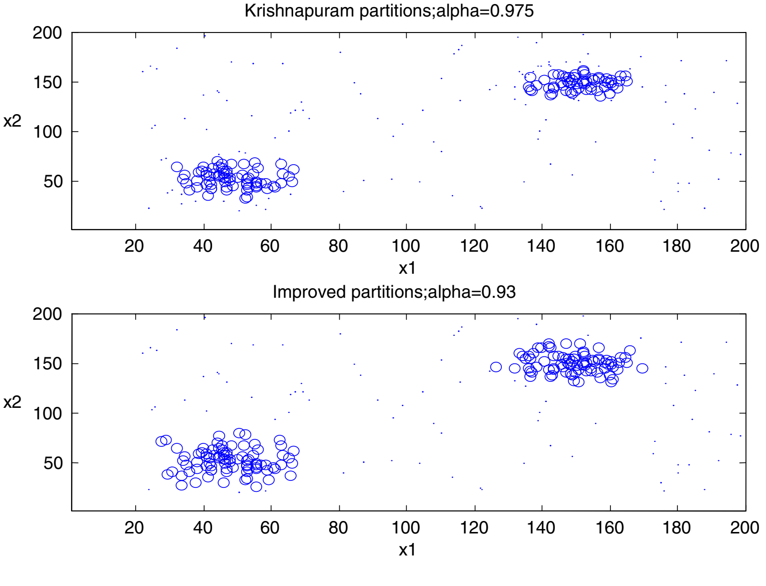

Fig. 5. Comparison between NPCM and the proposed clustering algorithm for example 5.1.

we are able to evaluate the partitions obtained by each clustering algorithm. In the NPCM partitions, the cluster centers are identified correctly. On the other hand, some of the inliers are identified as noise. That is due to the fact that these points are far enough from the centers of the clusters and so the algorithm rejects them. The improved algorithm correctly identifies the clusters and rejects the noise. The accuracy of the results obtained from the improved algorithm relies on the correct choice of the cutoff distance in Fig. 4.

5.2. Case study B

In this example, we generated a two-dimensional data set as shown in Fig. 6. The data was created manually in MATLAB 6.0 such that it forms three clusters and some random noise points. We have 100 data features including the noise. Fig. 7 shows the trace of the scatter matrix obtained from the fuzzy C-Means clustering algorithm. The suitable weight exponent for this example was found to be $m = 2.5$.

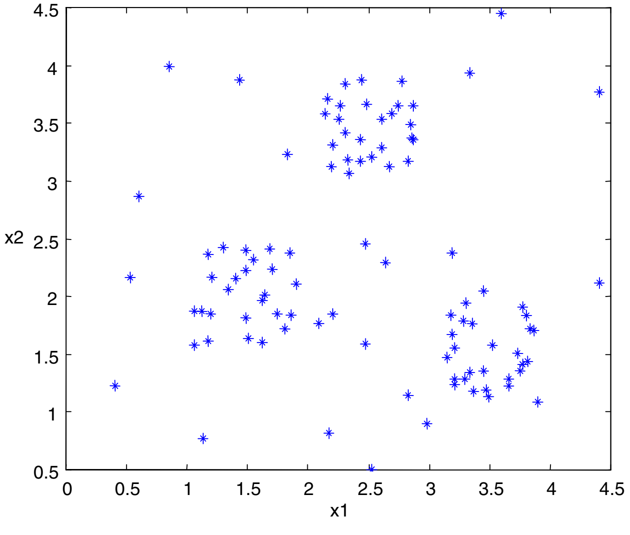

Fig. 6. The data set of case study B.

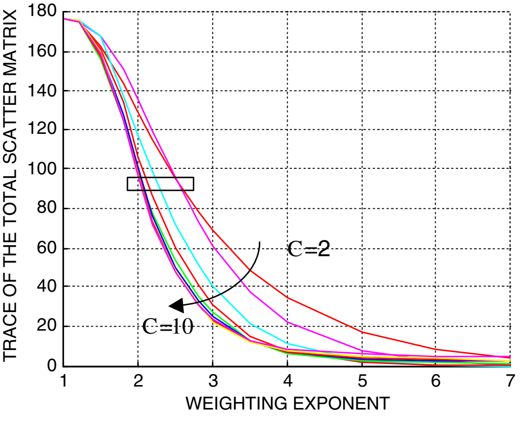

Fig. 7. Selection of the level of fuzziness for the example of case study B.

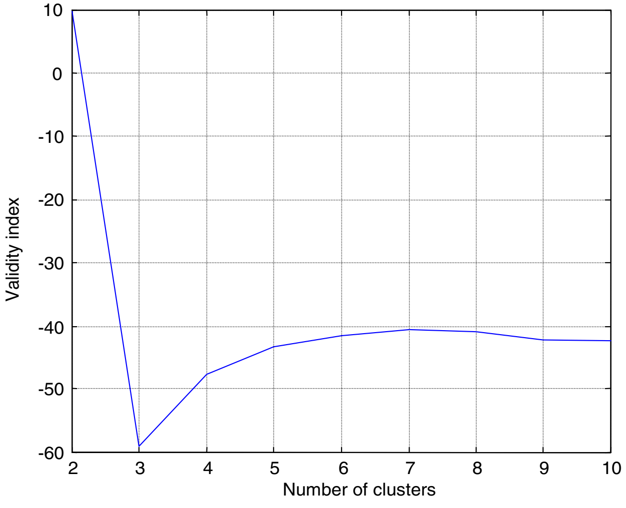

Fig. 8. Identification of the optimum number of clusters for the example of case study B.

Fig. 8 shows the validity index for this example. The optimum number of clusters is $c = 3$. At this stage, all the points are considered as inliers because the cluster validity index in (7) does not have a noise-rejection capability. The exact value of chi-square distribution to be used for this example can be assigned using Fig. 9. If we choose the cutoff distance at $\Omega = 6$, 12 data points are rejected. The cutoff distance to be used in the NPCM is calculated using $\chi^2_{0.88}$ instead of the $\chi^2_{0.975}$ suggested in [4]. Fig. 10 shows the comparison between the partitions obtained using the NPCM without the noise-

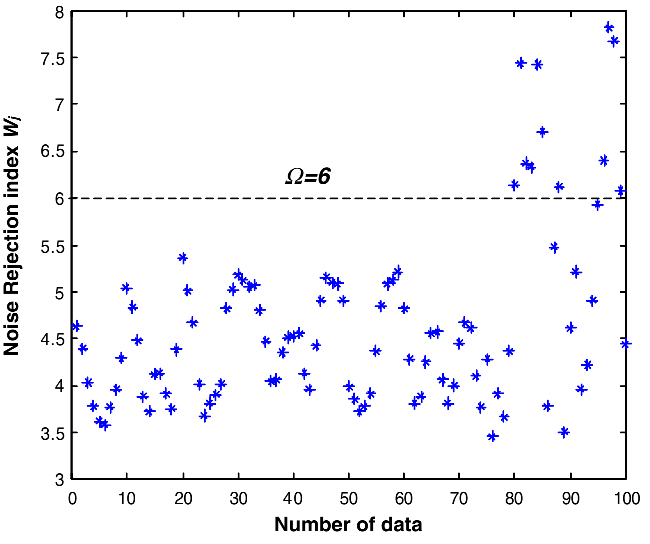

Fig. 9. Application of the noise-rejection criterion for the example of case study B.

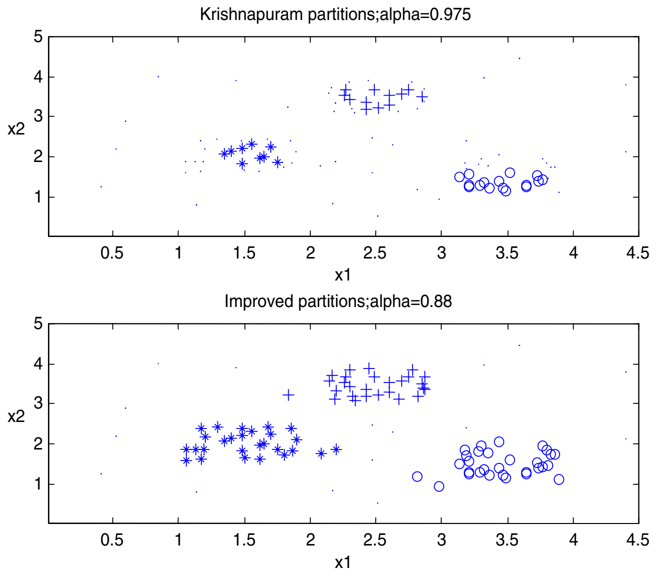

Fig. 10. Comparison between NPCM and proposed algorithm for case study 5.2.

rejection criteria and the partitions using the combined algorithm. The idea is to identify the clusters correctly as it is recognized by inspection where, as mentioned earlier, the noise points are known before applying the clustering algorithms. The

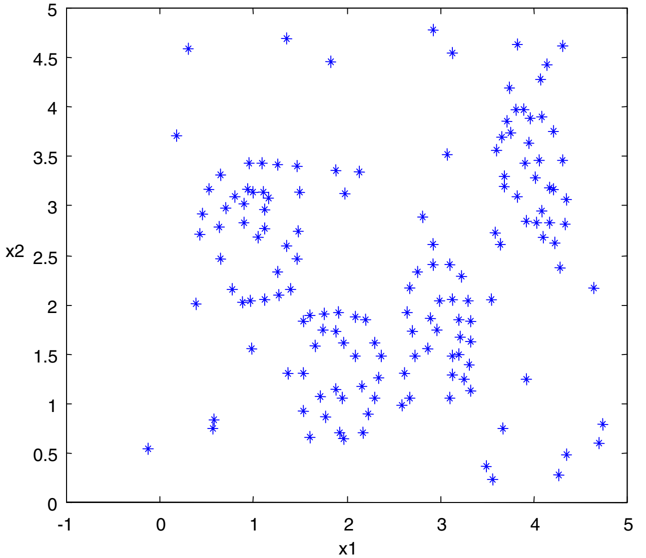

Fig. 11. The data set for the example of case study 5.3.

combined algorithm correctly identifies the cluster centers, rejects the noise points, and assigns the exact partitions in the data. Using the NPCM without the noise-rejection criterion would identify some of the points that belong to the clusters as noise.

5.3. Case study C

In this example, we generated another two-dimensional data set with 150 features including the noise. This example is designed to be more realistic having overlaps between the clusters. The data in Fig. 11 was also created in MATLAB 6.0 such that it forms four clusters and some random noise points. This example shows the capability of the combined algorithm to identify the exact data partitions in the

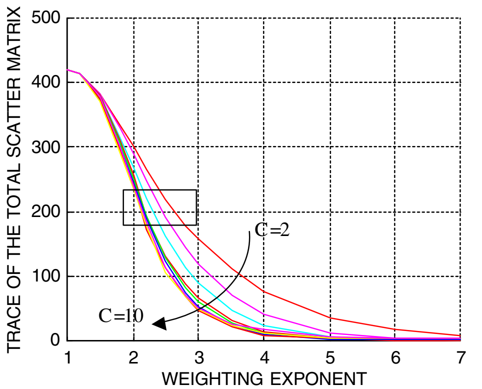

Fig. 12. Selection of the level of fuzziness for the example of case study C.

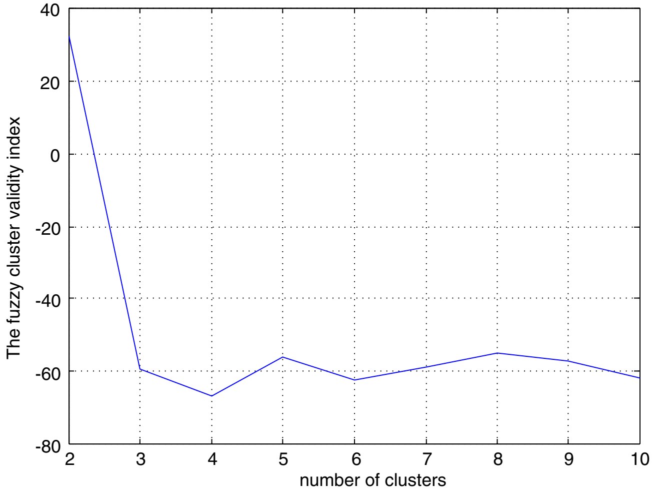

Fig. 13. Identification of the optimum number of clusters for the example of case study C.

presence of noise and outliers. Fig. 12 shows the trace of the total scatter matrix for different values of the weighting exponent. The suitable value for the weight exponent in this example is selected as $m = 2.5$. Fig. 13 shows the validity index for this example. The optimum number of clusters is $c = 4$. As shown in noise-rejection plot of Fig. 14, if the cutoff distance is chosen such that $\Omega = 10$, then 18 points on the curve are rejected. The cutoff distance in the combined algorithm will be calculated

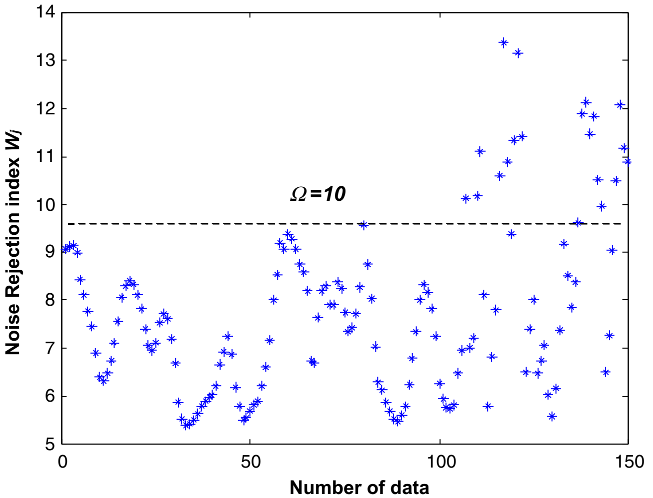

Fig. 14. Application of the noise-rejection criterion for the example of case study C.

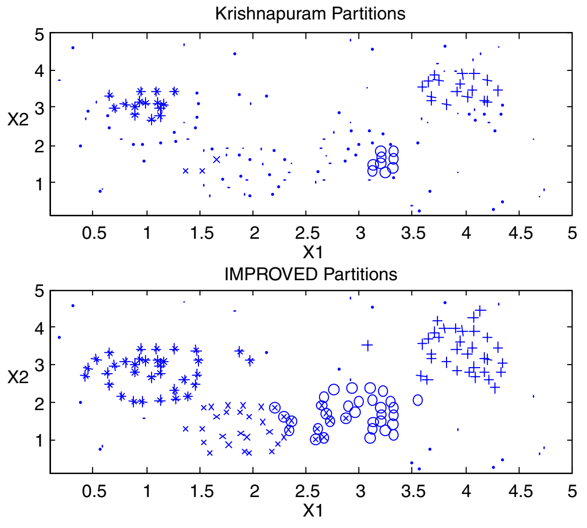

Fig. 15. Comparison between NPCM and proposed algorithm for case study 5.3.

using $\chi^2_{0.88}$. Fig. 15 shows a comparison between the NPCM partitions and the partitions obtained using the improved algorithm. The improved algorithm identified some of the inliers that were rejected in the NPCM algorithm. Moreover, both algorithms identify the cluster centers correctly in the presence of noise.

5.4. Case study D

In [5], the authors introduced the following nonlinear static system:

$$y = (1 + x_1^{-2} + x_2^{-1.5})^2; \quad 1 \leqslant x_1, \ x_2 \leqslant 5 \tag{16}$$

For the system in (16), 50 input–output data are randomly obtained. The output data points in Fig. 16 are to be clustered. Fig. 17 shows the trace of scatter matrix

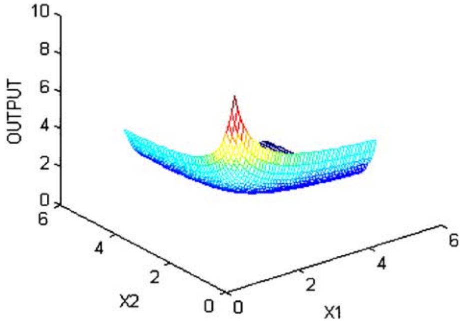

Fig. 16. Sugeno–Yasukawa static function of case study D.

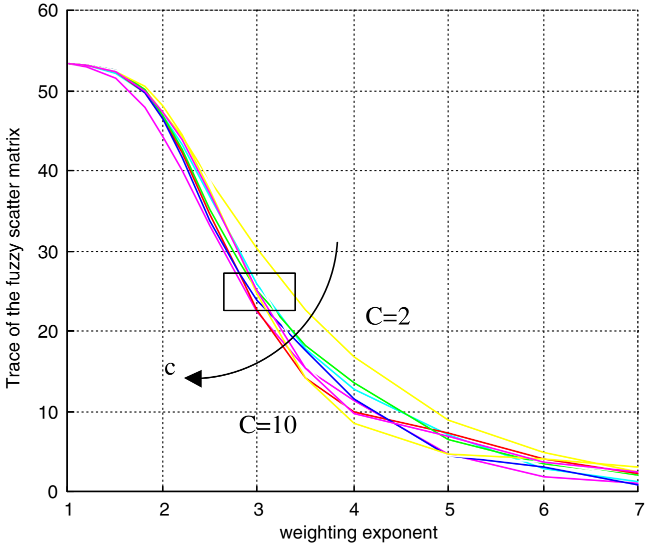

Fig. 17. Selection of the level of fuzziness for the example of case study D.

using the fuzzy C-Means clustering algorithm [7]. The suitable weight exponent for this example is $m = 3$. Fig. 18 shows the validity index with $c = 8$. Fig. 19 exemplifies the application of the noise-rejection criterion. For this specific example, it is difficult

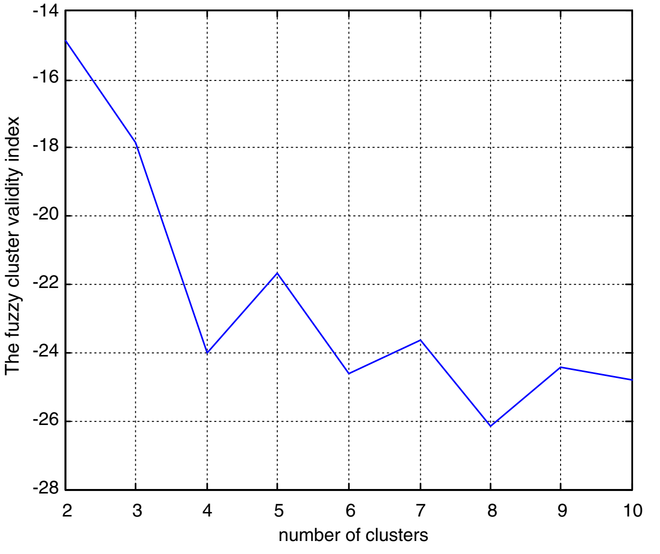

Fig. 18. Identification of the optimum number of clusters for the example of case study D.

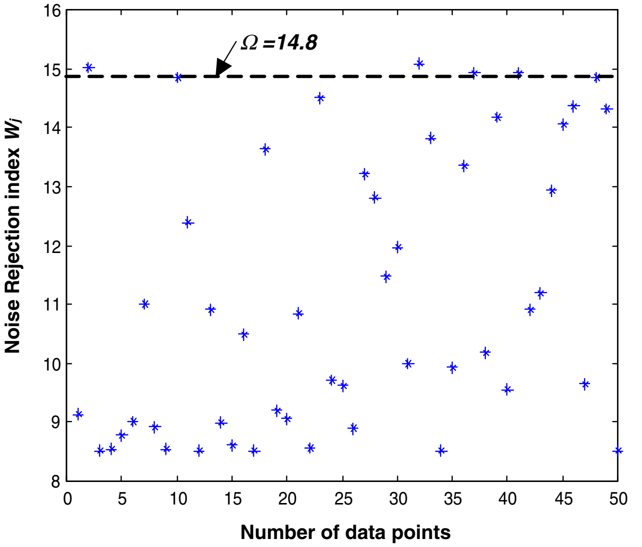

Fig. 19. Application of the noise-rejection criterion for the example of case study D.

to assign a cutoff distance. That is because the data is obtained from a deterministic function and so it is clean. In an attempt to find the exact partitions, we will choose the cutoff distance such that $\Omega = 14.8$. Hence, three points are rejected. The cutoff distance for the combined algorithm will be calculated using $\chi^2_{0.94}$.

6. Conclusions

In this paper, we reviewed the fuzzy C-Means clustering algorithm. The original FCM clustering algorithms has some weaknesses in identifying the initial cluster centers, selection of the weight exponent, and the assignment of the optimum number of clusters. In this paper, we introduced solution to treat the existing FCM problems related to the selection of the level of fuzziness, optimum number of clusters, and the locations of the clusters centers. On the other hand, the fuzzy C-Means clustering algorithms may fail completely in the presence of noise and outliers. Therefore, in [4] the authors introduced an improved Possibilistic C-Means algorithm for noise rejection. The improved Possibilistic clustering algorithm (NPCM) efficiently identifies the cluster centers in the presence of noise and outliers. However, the NPCM suffers from the same drawbacks of the original FCM clustering algorithms and introduces another parameter that must be identified a priori, i.e., the percentage of inliers in the data set. Therefore, in this paper we introduce a robust noise-rejection clustering algorithm. Unlike the NPCM, the proposed algorithm is based on a defined criterion for the assignment of the cutoff distance from the data. The matrix of dissimilarities is calculated for the data set and the noise-rejection curve is considered. This methodology is applicable for n-dimensional data set because the noise-rejection curve is always two-dimensional regardless of the dimension of the data features, i.e., only the two-dimensional data are observable. That is why the partitions obtained using the combined algorithm are more effective than those obtained by the NPCM.

References

[1] J.H. Ward, Hierarchical grouping to optimize an objective function, Journal of American Statistics Association (58) (1963) 236–244.

[2] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.

[3] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, NY, 1981.

[4] R. Krishnapuram, Generation of membership functions via possibilistic clustering, IEEE International Conference on Fuzzy Systems 2 (1994) 902–908.

[5] M. Sugeno, T. Yasukawa, A fuzzy-logic-based approach to qualitative modeling, IEEE Transactions on Fuzzy Systems 1 (1) (1993).

[6] Y. Fukuyama, M. Sugeno, A new method for choosing the number of clusters for the fuzzy C-Means method, in: Proceeding of the 5th Fuzzy Systems Symposium in Japanese, 1989, pp. 247–250.

[7] M.R. Emami, I.B. Turksen, A.A. Goldenberg, Development of a systematic methodology of fuzzy logic modeling, IEEE Transactions on Fuzzy Systems 6 (3) (1998) 346–361.

[8] J.L. Myers, A.D. Well, Research Design and Statistical Analysis, Harper Collins Publishers, 1991, pp. 204–233.
[9] W.W. Melek, Neurofuzzy Control of Modular and Reconfigurable Robots, PhD. Dissertation, University of Toronto, Canada, 2002.